HMW 1: due February 21, 4pm, upload on Canvas

Note: February 24 is a holiday, therefore I shortened the homework and moved the deadline to February 21.

Instructions:

- submit ONE PDF file containing the entire homework; we don't accept multiple files, even when uploaded in a ZIP folder
- include the name of each group member and your student numbers into the PDF
- include in the PDF the code you ran (if you use R Markdown it will be easy to do so).

Question 1

You are interested in finding patterns in pollution data. The data contains quantities in tonnes of different type of pollutants for different countries in the EU over several years. It is is collected by EEA and more explanations can be found on the EUROSTAT webpage.
https://ec.europa.eu/eurostat/cache/metadata/en/env_air_emis_esms.htm

Focus for now on the sulphur oxides data.

a) Run a PCA on the standardized observations using **prcomp()**, and display the first 2 PC loadings. Plot the first two PC loadings and interpret the plot, explaining also the statistical pitfalls in trying to interpret it. Explain what you are trying to achieve when running PCA (here the unit of observation is years, and the variables are air pollution for a given country).

b) Display a screeplot of the eigenvalues ordered from largest to smallest. Explain how many PC you would choose based on this screeplot, but also explain why such a conclusion may not be reliable.

c) Use instead the following BIC criterion: for k principal components, $BIC(k) = log(SSR(k)) + k\,\frac{log(np)}{np}$, to select the number of PC. How many components does the BIC select? *(Explanation: The number of variables p and the number of observations n is in this case equal, in*

which case the effective sample size C_{np} is not n, the number of time series observations, and not $min(n, p)$, but $np = 28^2$. Set $k < p$, else you get a perfect fit, and the BIC criterion does not work.

For the rest of the analysis, keep always the first two PC only. Redo the analysis for each type of pollutant in the data set.

d) Plot in one plot all the first PC over time for different pollutants (emissions). Do the same for the second principal component. Which pollutant trends look more similar over time? Explain.

Question 2

Explain, in a similar fashion to what we did in class on the blackboard, the intuition why the NIPALS algorithm is consistently estimating the PC solution to a factor model.

Question 3

Consider a two-class framework where we predict $Y \in \{1, 2\}$ with X, a scalar predictor. The prior class probabilities for Y are π_1 and π_2. Let $p_1(x) = P(Y = 1|X = x)$ and $p_2(x) = 1 - p_1(x) = P(Y = 2|X = x)$ be the probabilities that observation $X = x$ belongs to class 1 and class 2, respectively. Assume that in each class $k = 1, 2$, $X \sim \mathcal{N}(\mu_k, \sigma^2)$, and that we use the LDA classifier.

a) Write down the formula for the posterior probability $p_1(x)$ using the Bayes rule and explain how you obtained it.

Hint: the normal density of X in class k is $f_k(X = x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x - \mu_k)^2)$, and the Bayes' rule is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

b) Using a), prove that $\log \frac{p_1(x)}{1-p_1(x)} = c_0 + c_1 x$ by deriving the formulae for c_0, c_1 as functions of $\mu_1, \mu_2, \sigma^2, \pi_1, \pi_2$.

c) Explain what this derivation in b) shows about the relationship between LDA and logistic regression.

Consider now a two-class framework where we predict $Y \in \{1, 2\}$ with X, a vector of $p > 1$ predictors. In each class $k = 1, 2$, $X_k \sim N(\mu_k, \Sigma)$. The prior class probabilities for Y are π_1 and π_2. Assume that the density of X in class k is normal, meaning

$$f_k(X = x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)'\Sigma^{-1}(x - \mu_k)\right),$$

where $|\Sigma|$ is the determinant of the matrix Σ.

d) Using this along with Bayes' theorem, show that the Bayes classifier assigns an observation $X = x$ to class 1 if $\delta_1(x) > \delta_2(x)$, where $\delta_k(x) = x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log \pi_k$ for $k = 1, 2$.

Hint: the Bayes classifier assigns an observation $X = x$ to the class k for which the posterior probability $P(Y = k|X = x)$ is the largest.